Prospective, Longitudinal Assessment of Developmental Neurotoxicity

Joseph L. Jacobson and Sandra W. Jacobson

Psychology Department, Wayne State University, Detroit, Michigan

Methodological issues in the design of prospective, longitudinal studies of developmental neurotoxicity in humans are reviewed. A comprehensive assessment of potential confounding influences is important in these studies because inadequate assessment of confounders can threaten the validity of causal inferences drawn from the data. Potential confounders typically include demographic background variables, alcohol and smoking during pregnancy, the quality of parental stimulation, the child's age at test, and the examiner. Exposure to other substances is assessed where significant exposure is expected in the target population. In most studies, control variables even weakly related to outcome are included in all multivariate statistical analyses, and a toxic effect is inferred only if the effect of exposure is significant after controlling for the potential confounders. Once a neurotoxic effect has been identified, suspected mediating variables may be added to the analysis to examine underlying processes or mechanisms through which the exposure may impact on developmental outcome. Individual differences in vulnerability may be examined in terms of either an additive compensatory model or a synergistic "risk and resilience" approach. Failure to detect real effects (Type II error) is of particular concern in these studies because public policy considerations make it likely that negative findings will be interpreted to mean that the exposure is safe. Important sources of Type II error include inadequate representation of highly exposed individuals, overcontrol for confounders, and inappropriate correction for multiple comparisons. Given the high cost and complexity of prospective, longitudinal investigations, cross-sectional pilot studies focusing on highly exposed individuals can be valuable for the initial identification of salient domains of impairment. — Environ Health Perspect 104(Suppl 2):275–283 (1996)

Key words: prospective studies, longitudinal studies, neurotoxicity, prenatal exposure, neurobehavioral teratology, confounding variables, mediating variables, Type II error, risk and resilience, control variables

Introduction

The first studies to link intrapartum chemical exposure to behavioral deficits in the absence of organic damage were experiments with laboratory animals. Pioneering studies on hypervitaminosis A (1) and methylmercury (2) led behavioral teratologists to hypothesize that agents which produce mental retardation and severe neurological dysfunction at high doses will be associated with subtle behavioral changes when exposure occurs at lower levels (3). Animal studies in which dose and timing of exposure can be manipulated experimentally afford firmer causal inferences than human studies, where exposure may be confounded with extraneous variables that make its effects difficult to isolate. Moreover, the short lifespan of most laboratory species makes it possible to track long-term effects of perinatal insult that may not become evident in the human for several years (4,5).

Most human behavioral teratology studies have used prospective designs in which subjects are recruited prenatally or at birth and followed longitudinally. The principal advantages of a prospective design are more accurate assessment of degree of exposure, information regarding the timing of exposure, and more adequate assessment of relevant extraneous variables. For some substances, such as lead, exposure can be documented retrospectively in deciduous teeth (6) or in bone scans. For other substances such as cocaine or opiates, however, urine, meconium, or hair samples must be obtained contemporaneously; and for exposures such as alcohol, for which no reliable bioassays yet exist, self-report data must be obtained as soon as possible after exposure to limit memory decay. Even with lead, prospective ascertainment is necessary to determine the timing of exposure, which can be important both for investigating the mechanism of action and for devising intervention strategies. Extraneous variables, such as perinatal exposure to other contaminants and quality of intellectual stimulation provided by the parent, are often difficult, if not impossible, to assess retrospectively.

All developmental neurotoxicity studies using prospective, longitudinal designs have recognized the importance of assessing and controlling for a broad range of potential confounding influences. Investigators have differed, however, in their selection of control variables and in their strategies for identifying which potential confounders need to be included in multivariate analyses. Developmental neurotoxicity studies differ from many other longitudinal studies in that, in addition to the risk of spuriously attributing an observed effect to prenatal exposure (Type I error), failure to detect a real effect (Type II error) is also of particular concern. Despite our caveats that no inference should be made from a null finding, the need by policy makers and the general public to evaluate the risks associated with a potentially toxic exposure will inevitably lead negative findings to be interpreted to mean that the exposure is safe. Thus, a failure to detect real risks associated with an exposure may prevent necessary public health precautions and warnings from being implemented. Given the gravity of this risk, a power analysis (7) is usually required to

establish that the sample size is adequate to detect the real effects of the exposure.

This paper will address several methodological issues in the design of prospective, longitudinal studies of developmental neurotoxicity. We focus first on potential confounders, including criteria for their selection, alternative approaches to measurement, and strategies for selection for inclusion in multivariate analysis. The statistical treatment of mediating variables will also be considered, along with strategies for evaluating factors that may either enhance vulnerability or protect against the harmful effects of a developmental neurotoxic exposure. We will then review several factors that can increase the risk of Type II error, including inadequate representation of highly exposed individuals, overcontrol for confounders, and inappropriate correction for multiple comparisons. Finally, we will consider the degree to which, despite their limitations, retrospective studies may be useful in supplementing what can be learned from prospective, longitudinal investigations.

Control for Potential Confounders

Criteria for Selection

The selection of control variables to test for spurious correlation starts with the premise that an extraneous variable cannot be the true cause of an observed relation between toxic exposure and developmental outcome unless it is related to both exposure and outcome (8). In most studies relation to outcome is used as the criterion to select control variables, probably because more is usually known about the determinants of the developmental outcome than about the correlates of the exposure. Where physical growth is the focus, height and weight of both parents and child's sex are important determinants; where cognitive competence is of interest, it is important to assess the quality of intellectual stimulation and emotional support provided by the parents. Both sets of outcomes could be affected by perinatal risk variables (e.g., neonatal asphyxia) and other exposures, such as to alcohol, which has been linked to both growth retardation and cognitive deficit.

It is important that the measures selected to represent the potential confounders be both reliable and valid because inadequate measurement can threaten the validity of any causal inferences drawn from the data. Whereas unreliable measurement of exposure or outcome will increase the risk of failure to detect a real effect, inadequate

measurement of a potential confounder will tend to underestimate its influence on the outcome, possibly leading to the erroneous attribution of an observed effect to the exposure. For this reason, it is generally desirable to use standard measures with demonstrated validity for control variable purposes. Where new measures are constructed, it is important to check their convergent validity in terms of whether they correlate as predicted with outcome or other control variables (9). For example, if quality of parental supervision is considered an important influence on school achievement in dangerous, disorganized inner-city neighborhoods (S Hans, personal communication), a parental supervision scale could be constructed and validated in relation to academic achievement and the HOME Inventory.

Measurement of Potential Confounders

Table 1 provides a list of control variables that have been used in developmental neurotoxicity studies. At a minimum, most contemporary studies assess the demographic background variables listed in the table, alcohol and smoking during pregnancy, the quality of parental stimulation (usually the HOME Inventory), the child's age at test, and the examiner. Pregnancy alcohol and smoking are usually included because they are so prevalent; exposure to other substances would be assessed if there were reason to expect significant exposure in the target population. Although for most substances intrauterine exposure seems to pose the greatest threat, postnatal exposure may also be relevant. Breast-feeding exposure, which can be significant for polychlorinated biphenyls (PCBs), organochlorine pesticides, and other lipophilic substances, is assessed in terms of two variables: contaminant levels in maternal milk and amount of contaminated milk consumed. The latter is indicated most reliably by duration of breast-feeding (21). Postnatal environmental exposure to lead (e.g., from paint chips or dust) can be assessed by obtaining serial blood lead levels from the child (22,23). Some recent studies have incorporated increasingly detailed assessments of socioenvironmental influences in light of contemporary risk and resilience models suggesting that the long-term functional effects of an initial teratological insult may depend in some cases on the presence of co-morbid environmental risk factors (24–26). Examiner can be used to adjust child test scores for subtle differences in test administration by different examiners.

In contrast to smoking during pregnancy, which has high test–retest reliability even over a period of several years (9), alcohol and drug use are difficult to recall reliably and are often highly stigmatized. Some studies have used a dichotomous yes/no measure to summarize maternal drinking during pregnancy. Given what is known about the teratological effects of alcohol, however, a use-versus-abstinence measure cannot adequately control for alcohol exposure. Because most women drink less than 0.5 oz absolute alcohol per day (AA/day), the lowest level at which effects are typically seen (27), grouping a large number of light drinkers together with the relatively small number whose drinking puts their infants at serious risk is likely to obscure the true effects of the prenatal alcohol exposure in the analysis and to underestimate the effects of the alcohol exposure for control variable purposes.

Table 1. Control variables for consideration in developmental neurotoxicity studies.

Demographic background
  Socioeconomic status (e.g., Hollingshead, unpublished)
  Race
  Marital status
  Mother's age
  Parity and/or number of children
  Child's sex

Perinatal risk
  Perinatal Risk Scale [e.g., Molfese et al. (10)]
  Number of prenatal clinic visits
  Gravidity

Neurotoxic exposures
  Alcohol
  Smoking
  Illicit drugs (e.g., cocaine)
  Lead
  Methylmercury

Socioenvironmental influences
  Quality of parental intellectual stimulation [e.g.,
    HOME Inventory (11); Nursing Child Assessment
    Teaching Scales (12)]
  Parental intelligence [e.g., Peabody Picture
    Vocabulary Test–Revised (13)]
  Nursery school attendance or elementary school
    experience
  Quality of school
  Family stress [e.g., Life Events Scale (14)]
  Maternal social support [e.g., Crnic et al. (15)]
  Maternal depression [e.g., Beck Depression
    Inventory (16)]
  Maternal psychopathology [e.g., SCL-90R (17);
    Personality Diagnostic Questionnaire–Revised (18)]
  Family cohesiveness [e.g., Family Environment
    Scale (19)]
  Marital conflict [e.g., Conflict Tactics Scales (20)]

Situational
  Child's age at test
  Medications within past 4 to 8 hr
  Examiner

The standard approach to quantifying maternal drinking during pregnancy is a quantity–frequency–variability (Q-F-V) interview (28) in which the mother is asked how much she drinks on the days she consumes alcohol, how many days per week she drinks, and how much and how often she drinks at higher and lower levels. This information is obtained separately for beer, wine, and liquor, and volume is converted to ounces of (AA) based on the alcohol content of the beverages consumed (29). One drink of beer, wine, or liquor is equivalent to approximately 0.5 oz of AA. Among the summary variables that can be constructed from these data, oz AA/day averaged across pregnancy has proven the strongest. Other summary measures include proportion of pregnancy days when drinking occurred, average AA per drinking day (volume/occasion), and bingeing (e.g., whether the mother drank at least 2.5 oz AA [5 standard drinks]) on one or more occasions during the index pregnancy). Our research indicates that a summary measure based on multiple self-reports obtained periodically during pregnancy is markedly more reliable than a single maternal interview (30).

Intrapartum use of illicit drugs, such as cocaine, opiates, and marijuana, can now be ascertained by biological assay of meconium or maternal urine or hair. Biological assays are critical, given the high rate of denial associated with maternal self-reporting of illicit drug use (31). Zuckerman et al. (32) found effects of cocaine exposure on birth size using a dichotomous use-versus- abstinence measure based on evidence from either self-report or urine assay but not on a use/abstinence measure based solely on self-report. Cocaine is detectable in urine samples for 3 days (33), in meconium for up to 6 months (34,35), and in hair for several months depending on length (hair grows at a rate of approximately 1 cm per month) (1). The principal disadvantage of the assays currently available is that they provide no information on degree of exposure. Since, as with alcohol, risk to the fetus may be associated only with moderate-to-heavy drug use, it is important to supplement biological assays with self-report data obtained during pregnancy. Although a comprehensive Q-F-V approach can be used (36), the quantity dimension is likely to be unreliable due to the wide variability in the dosage and degree of purity of illicit street drugs. Once exposure has been determined by biological assay, self-report frequency data may be sufficient to discriminate moderate and heavy from lighter users.

Measures of socioeconomic status (SES) based on parental education and occupational status (AB Hollingshead, unpublished) explain considerable variance in child cognitive performance (37), presumably because better educated, higher SES parents tend to provide more optimal intellectual stimulation to their children. Because SES is only an indirect indicator of the quality of parental input, however, instruments such as the HOME Inventory (11) have been developed to provide a more direct assessment. The HOME combines a semistructured interview with informal observation of parent–child interaction to evaluate the quality of intellectual stimulation and emotional responsiveness provided by the parent. Caldwell and Bradley (11) recommend that the information required for the HOME Inventory protocol be elicited informally and spontaneously from the mother. S.W. Jacobson (unpublished) has prepared scripts for the infant, preschool, and elementary school versions of the HOME, based on the probes suggested by Caldwell and Bradley, which reorganize and standardize the presentation of the interview material to facilitate this approach. Three versions of the HOME are available—infant through age 3 years; preschool, 3 to 6 years; and elementary school, 6 to 10 years. The HOME provides a more comprehensive assessment of parental input than SES: data show that it explains significant variance in cognitive performance over and above standard SES measures (38–40). Although designed to be administered in the home, the assessment can be modified for use in the laboratory when logistical considerations preclude home visits (41).

Although listed under socioenvironmental influences in Table 1, parental intelligence influences child cognitive performance through genetic endowment as well as quality of intellectual stimulation. Statistical control of both these sources of influence is frequently warranted in a teratological study since both are extraneous to the teratological process under investigation. Because it is rarely feasible to perform a full IQ test on parents and because vocabulary is the strongest single correlate of IQ, the Peabody Picture Vocabulary Test–Revised (PPVT-R) (13) is often used to assess parental intelligence for control variable purposes. The PPVT-R is strongly correlated with standardized tests of adult IQ, and, although minority subjects tend to score low due to limited educational opportunity, the test has been shown to be

valid for rank ordering lower SES, black mothers within a homogeneously disadvantaged sample (42). Additional dimensions of socioenvironmental influence that may warrant consideration in studies of cognitive performance include nursery school attendance, months of experience in formal classroom settings, and quality of school attended (e.g., inner city, urban magnet, parochial, private, etc.).

The HOME Inventory, parental intelligence, and formal school experience provide a comprehensive assessment of socioenvironmental influences on intellectual development, but other control variables may be more relevant where the focus is social and affective development. For example, it has been suggested that prenatal cocaine exposure may impact strongly on emotional arousal and motivation (43), and nonretarded, fetal alcohol syndrome adults have been described as exhibiting poor judgment and an inability to respond to subtle social cues (44). Because less is known about socioenvironmental influences on social and affective development, a broader range of control variables warrant consideration. Examples listed in Table 1 include familial stress, maternal social support, maternal depression and psychopathology, family cohesiveness, and marital conflict.

Selection for Inclusion in Multivariate Analysis

Multivariate analysis is used to determine the degree to which effects of exposure are seen after statistically removing the influence of potential confounders. Although some researchers (45) have advocated including all control variables in every analysis, that approach has at least two disadvantages. Where a large number of control variables are included, the coefficient assessing the magnitude of the toxic effect is likely to be unreliable; a minimum of 20 subjects per independent variable is recommended (46). In addition, the inclusion of control variables unrelated to the outcome will tend to increase the size of the error term, making it more difficult to detect significant toxic effects (47). For these reasons, we have adopted the procedure of prescreening the control variables to decide which to include in multivariate analyses.

As with the determination of which control variables to assess, the selection of potential confounders for inclusion in the statistical analyses is based on the premise that a control variable cannot be the true cause of an observed effect of exposure on outcome unless it is related to both (8). In

our research on the effects of prenatal PCB exposure (48), control variables were selected for inclusion based on their relation to exposure. Any control variable related to an exposure measure (at $p < 0.10$) was included as a potential confounder in all analyses evaluating the effects of that exposure. In our more recent research on prenatal alcohol exposure, however, control variables were selected in relation to outcome rather than exposure (30,49). Selection in relation to outcome is preferable because, where a control variable unrelated to exposure explains some variance in the outcome, its inclusion reduces the error term, thereby improving the chances of detecting toxic effects (47). Relation to outcome is the criterion used most commonly in contemporary developmental neurotoxicity studies (50–52). Control variables are typically included if they are associated with outcome at $p < 0.10$, which is conservative in this context because it includes even weak potential confounders in the analysis. A toxic effect is inferred only if the relation between exposure and outcome is significant at $p < 0.05$ after controlling for the potential confounders.

A different approach, recommended by Kleinbaum et al. (47), involves the initial entry of all control variables in the analysis, followed by stepwise removal of all variables whose deletion does not substantially alter the magnitude or precision of the effect of exposure in the analysis. In multiple regression, magnitude refers to the size of the standardized regression coefficient associated with exposure; precision refers to its confidence interval or statistical significance. In principle, this approach is sound since only those potential confounders whose inclusion alters the relation between exposure and outcome are relevant for statistical control purposes. Kleinbaum et al. (47) recommend that the investigator retain in the analysis only those confounders whose removal alters the effect on outcome sufficiently to be considered clinically important. Unfortunately, this approach is difficult to implement because there is little consensus among investigators regarding what magnitudes are functionally significant.

Mediating Variables

Once a teratogenic effect has been identified, the focus shifts to an examination of the underlying processes or mechanisms through which the neurotoxic exposure impacts on the outcome. For example, the effect of prenatal cocaine exposure on birth weight has been explained in terms of

cocaine's action as an appetite suppressant (53) and as a vasoconstrictor (54). The vasoconstriction hypothesis is based on experiments with sheep showing that cocaine-induced vasoconstriction decreases uterine blood flow, thereby limiting transfer of nutrients and oxygen to the fetus (55). Appetite suppression and vasoconstriction are considered mediating or intervening variables in these explanations. There is also considerable interest in socioenvironmental mediating variables. O'Connor et al. (56) have shown, for example, that the effect of prenatal alcohol exposure on the Bayley Mental Development Index (MDI) at 1 year of age is mediated, in part, by temperamental irritability in alcohol-exposed infants, who do poorly on the Bayley because they elicit less optimal intellectual stimulation from the parent. Hypotheses incorporating mediating variables are tested most effectively by structural equation modeling procedures, such as LISREL (57).

Although relevant potential confounders should be included in all statistical analyses, the routine inclusion of mediating variables can be misleading. Confusion can arise because confounders and mediators are tested statistically in the same manner. For example, an effect of prenatal cocaine exposure on neurobehavioral outcome could be mediated by reduced birth size. Mediation can be tested by adding birth size to the analysis of the cocaine effect on neurobehavior; if the cocaine effect is no longer significant, mediation by birth size is inferred. If birth size were a confounder and its inclusion rendered an observed cocaine effect nonsignificant, one would conclude that the cocaine effect was spurious. But where birth size is a consequence of the exposure, mediation is the appropriate interpretation. Potential confounders should be included routinely in all analyses because effects of exposure are of interest only after alternative explanatory variables

have been statistically controlled. Mediators should not be entered in the initial analyses evaluating toxic effects, however, because their effects can be understood only if analyses excluding them are compared with analyses that include them.

Vulnerability and Protection

Until recently, most developmental neurotoxicity studies have been premised on a biologically based main effects model in which organic damage early in development is assumed to lead directly to childhood cognitive or behavioral deficits. More recent studies have begun to consider the alternative view that subtle deficits may result from an interaction between an initial insult and co-morbid biological or environmental factors that may be necessary to sustain the initial teratological damage or that contribute to its emergence (24–26). Contemporary risk and resilience models were originally formulated in studies of the offspring of mentally ill parents to explain why many children seemed to escape relatively unscathed. Marked variability also characterizes the findings in developmental neurotoxicity studies. For example, Table 2 shows that children prenatally exposed to PCBs at relatively high levels are more than twice as likely to exhibit poor performance on the McCarthy Memory Scales at 4 years of age. Nevertheless, 12 of the highest exposed children performed in the normal range and 1 performed exceptionally well. Individual differences in vulnerability are not limited to the relatively subtle deficits seen at the moderate exposure levels in our PCB research. A large proportion of infants exposed prenatally to high levels of alcohol fail to develop fetal alcohol syndrome (58), and, even among those who do, many exhibit normal range IQs (44).

Individual differences in vulnerability can be explained in terms of a compensatory model. The parents of the exceptionally

Table 2. McCarthy Memory Scale performance by prenatal PCB exposure.

	Prenatal PCB exposure			
	Low	Moderate	High	Total
Memory scale[a]				
More than 1 SD below mean	14 (13.6%)	9 (19.6%)	8 (38.1%)	31 (18.2%)
Within 1 SD of mean	72 (69.9%)	35 (76.1%)	12 (57.1%)	119 (70.0%)
More than 1 SD above mean	17 (16.5%)	2 (4.3%)	1 (4.8%)	20 (11.8%)
Total	103	46	21	170

Values are cell frequencies. Abbreviations: SD, standard deviation; high, cord serum PCB level >5.0 ng/ml or maternal milk level >1250 ng/ml; moderate, cord serum PCB level = 3.0–4.9 ng/ml or maternal milk level = 1000–1,249 ng/ml; low, low on both cord and maternal milk. $\chi^2(4) = 11.54$, $p < 0.05$. [a]After adjustment for the potential confounders of the cord serum and breast milk PCB measures: maternal age, gravidity, examiner, maternal employment, alcohol consumption before pregnancy, breast milk PBB level, and 4-year blood lead level.

performing, high-exposed child in Table 2 may have worked intensively with him or her to overcome the limitations imposed by an organically based deficit. Statistically, compensation posits an additive model since high quality parental input is seen as reducing the severity of the deficit. By contrast, Rutter's (26) resilience model posits statistical interaction. Neurotoxic exposure constitutes a risk whose consequences may depend on one or more factors that may render the individual vulnerable or resilient. In a study of institution-reared women, Rutter and Quinton (59) found that depressed mothers were more likely to target children with difficult (irritable, moody) temperaments as outlets for excessive hostility. The data suggest a synergistic rather than an additive model. When children of depressed mothers had easygoing or average temperaments, they experienced very low levels of parental hostility. Easy or average temperament did not reduce the level of parental hostility; it precluded the child's becoming the target.

Differential vulnerability to a neurotoxic agent may also be attributable to differences in the timing of the exposure (critical or sensitive period) or to individual differences in genetic makeup or metabolism. Jacobson et al. (30) found alcohol-related deficits on the Bayley Scales only in the offspring of mothers over 30 years of age, suggesting that vulnerability may depend on physiological changes in the mother associated with a history of heavy drinking.

By contrast to models incorporating mediating variables, which can be tested by adding continuous measures to a multiple regression or structural equation model analysis, the risk and resilience approach posits a statistical interaction. The vulnerability or protection factor is considered a moderator variable, which cannot readily be incorporated in a structural equation model but can be tested by adding an interaction term to a multiple regression analysis. Unfortunately, the power of the significance test for a statistical interaction is low (7), in part because only a small proportion of the sample may be vulnerable or, conversely, protected. Extensive exploratory analyses may be necessary to identify the cut points at which vulnerability becomes operative to avoid grouping large numbers of nonvulnerable children together with the few truly at risk for the adverse outcome. Analysis is further complicated by Rutter's (26) observation that adverse effects are often seen only in the presence of two or more vulnerability factors.

Type II Error

Sampling from the Highest Exposed Individuals

Although Cohen's (7) power analysis is important for insuring that the study sample is large enough to detect the neurotoxic effects of a prenatal exposure, inadequate sample size is only one of several potential sources of Type II error. One of the most significant risks in a developmental neurotoxicity study involves the failure to oversample adequately from among the most highly exposed individuals. Although the prevalence of a given exposure is an important research focus for the epidemiologist, the first priority of the behavioral teratologist is to ascertain any deleterious effects and, if any are found, to assess their severity. Oversampling from the highest exposed individuals is critical because, if there are effects, these children will be the most likely to reveal them and to exhibit the most severe impairment.

The importance of oversampling became clear to us upon reviewing the literature on the effects of prenatal alcohol exposure on the Bayley Scales. Although Streissguth et al. (60), our group (30), and others (e.g., Smith et al., unpublished data) found effects on the Bayley, two major studies—one in Cleveland (61) and the other in Pittsburgh (62)—did not. In analyzing our data, we performed a contingency table analysis in which the bottom tenth percentile of the distribution was used to evaluate the incidence of poor performance on the Bayley MDI. This analysis showed an increased incidence of poor performance above a threshold of 0.5 oz AA/day during pregnancy (Table 3). An examination of the Cleveland data revealed that their sample included only 7 infants whose mothers drank above that threshold, compared with 45 in our sample, suggesting that their cohort contained too few infants exposed in the range in which the MDI effect is clearly seen. Moreover, when we randomly deleted all but 7 of the infants whose mothers drank

above the 0.5 oz threshold, the zero-order correlation of alcohol with the MDI dropped from –0.17 to –0.05, similar to the –0.06 correlation reported in Cleveland. If moderate-to-heavy drinkers had not been overrepresented in the other alcohol studies, the effects on the MDI would never have been detected.

Adequacy of sample size in terms of a Cohen (7) power analysis provides no assurance that high-exposed individuals have been adequately represented. Adequacy of representation can be determined only on the basis of data from previous studies indicating exposure thresholds above which effects are seen. Some oversampling was performed in the Pittsburgh study, but the criterion (>3 drinks per week) may have been too low to insure the inclusion of sufficient numbers of infants exposed above the 0.5 oz (7 drinks per week) threshold. Where no previous data exist, retrospective pilot studies may be warranted to suggest exposure levels above which effects might be expected.

Overcontrol for Confounders

A second potential source of Type II error in a developmental neurotoxicity study relates to routine control for potential confounders. This risk is illustrated by comparing data from two large prospective studies of the effects of lead exposure on childhood cognitive function. Bellinger et al. (22) studied low-level lead exposure (mean 24-month blood lead level = 6.8 µg/dl) in a predominantly white, college-educated, middle-class, suburban Boston sample. Dietrich and associates (51) studied somewhat higher level exposure (mean 24-month blood lead level = 17.0 µg/dl) in a predominantly black, poor, inner-city Cincinnati sample. The Boston study found that preschool-age blood lead level was associated with poorer performance on the McCarthy Perceptual Performance Scale, which indicated a significant visual–spatial deficit, after adjusting for 13 control variables including social class, maternal IQ,

Table 3. Number of infants scoring in the bottom tenth percentile[a] by pregnancy drinking level.

Drinking level	AA per day, oz	n	Bayley Mental Development Index [b]
Abstainer	0.00	60	4 (6.7)
Low	0.01–0.24	233	22 (9.4)
Low moderate	0.25–0.49	37	2 (5.4)
Moderate	0.50–0.99	26	6 (23.1)
Heavy	1.00–1.99	13	2 (15.4)
Very heavy	2.00+	6	1 (16.7)
Total sample		375	37 (9.9)

[a]After adjustment for potential confounders. [b]Values are number (percent) of infants at each level of exposure.

and the HOME Inventory. In Cincinnati, zero-order correlations indicated a relation between lead exposure and poorer performance on the Simultaneous Processing Scale of the Kaufman Assessment Battery for Children, which assesses the same domain as the McCarthy Perceptual Performance Scale. After controlling for only seven control variables, however, the lead effect was no longer significant. Hierarchical regression analysis showed that the lead effect remained significant until maternal IQ and the HOME were entered (Table 4).

The simplest interpretation of the data in Table 4 is that the zero-order correlation of lead with the Kaufman Scale is spurious and due to the fact that the lead-exposed children received poorer intellectual stimulation from their mothers. Alternatively, one might speculate that low SES may contribute to poorer cognitive performance by increasing the likelihood of a child's being raised in a dilapidated house containing lead-contaminated paint. If so, lead exposure may function as a mediating variable, that is, a mechanism whereby SES may influence cognitive performance. In Cincinnati, SES and lead exposure were apparently too highly confounded for a lead effect to be detected. Paradoxically, in Boston where the lead level was lower, the effect was easier to detect, either because quality of stimulation was unrelated to lead in the more middle-class sample or because most of the parents in that sample provided at least minimally adequate intellectual stimulation.

If only the 4-year Cincinnati lead data had been available, one might have erroneously concluded that lead has no effect at these levels of exposure. The evidence from Boston of an effect at even lower levels after control for confounding suggested the possibility that exposure and socioenvironmental influences may have been too confounded to separate statistically in Cincinnati. This suspicion was confirmed in a 6.5-year follow-up in Cincinnati,

Table 4. Effects of lead exposure on the Kaufman Assessment Battery for Children Simultaneous Processing Scale.

	r	β at Step 1[a]	β at Final Step[b]
36-month lead level	−0.26**	−0.20*	−0.11
48-month lead level	−0.31**	−0.26**	−0.15

*$p < 0.05$. **$p < 0.01$. [a]After adjustment for birth weight, maternal smoking during pregnancy, marijuana during pregnancy, race, and preschool attendance. [b]Also adjusted for HOME Inventory and maternal IQ.

which reported an effect of lead exposure on WISC-R Performance IQ after controlling for all relevant potential confounders (23). The authors attributed the 6.5-year finding to the greater reliability and precision of the WISC-R. Alternatively, 4-year test scores and social environment may be especially difficult to separate because performance by the 4-year-old, who has not yet attended school, depends so heavily on the quality of intellectual stimulation provided at home. Thus, although valid causal inference requires careful control for confounding, where exposure is highly confounded with an extraneous variable such as social environment, control for confounding can sometimes obscure potentially important causal effects.

One approach to reduce confounding would be to use developmental outcomes that are relatively insensitive to socioenvironmental influence. Table 5 shows the relation of the principal cognitive outcomes assessed in our PCB 4-year follow-up study (48) to selected socioenvironmental potential confounders. IQ (represented by the McCarthy General Cognitive Index) and child's vocabulary (PPVT-R) are much more strongly related to SES and quality of parental intellectual stimulation (HOME Inventory) than tests designed to focus more narrowly on short-term memory, visual discrimination, or attention. Although even performance on the vigilance task is influenced by quality of parental stimulation, the correlations are relatively modest, which may enhance the potential of this more narrow-band assessment to detect teratogenic effects. One of the principal advantages of assessment during the first postpartum year is that infant performance is relatively insensitive to sociocultural influences (see Table 6) (37,63). Even during infancy, narrow-band assessments, such

as the Fagan Visual Recognition Memory Test or infant reaction time (31,64), are less influenced by socioenvironmental factors than the more apical Bayley Scales.

Control for Multiple Comparisons

Where the specific effects of a prenatal exposure are not known in advance or deficits are suspected in multiple domains, the investigator may want to assess a large number of developmental outcomes. Given the high cost of recruiting and maintaining a prenatally exposed cohort and of assessing the necessary potential confounders, it makes sense to obtain as comprehensive a picture as possible of the nature of the impairment. However, a comprehensive test battery with a large number of outcome measures raises the concern that, where many outcomes are assessed simultaneously, a certain proportion will be significant by chance.

One traditional approach for dealing with multiple comparisons is the Bonferroni correction. Instead of using $p < 0.05$ as the criterion to reject the null hypothesis, 0.05 is divided by the number of outcomes assessed so that, if 20 outcomes are tested, a $p < 0.0025$ criterion would be used, making chance findings much less likely. The principal problem with the Bonferroni correction is an increased risk of Type II error. Reliable effects can easily be missed if all those between $p < 0.0025$ and 0.05 are considered nonsignificant. A better solution is to assess a broad range of outcomes in terms of the usual $p < 0.05$ criterion while recognizing that the use of multiple measures will increase the risk of Type I error in the short run. Inferences must be considered highly tentative if the number of significant effects seen does not exceed the number expected by chance. Even where multiple effects are seen, any unpredicted

Table 5. Relation of selected socioenvironmental potential confounders to 4-year cognitive outcomes.

	n	SES	HOME Inventory	PPVT-R	Mother's age	Nursery school attendance[a]
McCarthy General Cognitive Index	323	0.34***	0.50***	0.38***	0.11*	0.14**
Peabody Picture Vocabulary Test–Revised (child)	323	0.36***	0.46***	0.42***	0.18**	0.17**
Short-term memory for pictures (errors)	278	−0.10+	−0.17**	−0.02	−0.12+	−0.05
Visual discrimination						
Reaction time	279	0.09	0.09	0.12*	0.08	−0.01
Number correct	303	0.21***	0.39***	0.27***	0.06	0.15**
Vigilance (sustained attention)						
Reaction time	290	−0.02	−0.19**	−0.07	0.01	0.04
Number correct	290	0.11+	0.25***	0.08	0.04	0.06
Errors of commission (impulsivity)	290	−0.08	−0.18**	−0.02	0.04	−0.05

[a]Missing for one case. +$p < 0.10$. *$p < 0.05$. **$p < 0.01$. ***$p < 0.001$.

Table 6. Relation of selected potential confounders to infant cognitive outcomes.

	n	HOME Inventory[a]	Peabody Picture Vocabulary Test– Revised (mother)[a]	Mother's age	Maternal depression[a]	Number of prenatal clinic visits
Bayley Scales						
Mental Development Index	382	0.19***	–0.00	–0.21***	–0.10*	0.14**
Psychomotor Development Index	375	0.14**	0.02	–0.12*	–0.13*	0.11*
Fagan Visual Recognition Memory Test						
Novelty preference	315	0.04	–0.00	–0.02	–0.01	–0.02
Fixation duration	315	–0.12*	–0.10+	–0.00	0.12*	–0.04
Cross-modal Transfer Test						
Novelty preference	362	0.10+	0.04	–0.03	–0.02	–0.11*
Fixation duration	362	0.00	–0.04	0.07	0.04	–0.04
Complexity of play						
Spontaneous	304	0.02	0.07	–0.05	0.02	0.09
Examiner-elicited	310	0.04	0.04	–0.03	–0.04	0.05
Object permanence	354	0.08	0.16**	0.10+	–0.05	–0.06

[a]HOME Inventory is missing for 7 cases, Peabody for 17, maternal depression for 16. +$p < 0.10$. *$p < 0.05$. **$p < 0.01$. ***$p < 0.001$.

Table 7. Maternal recall reliabilities: Infancy to 4 years of age.

Birth weight	0.98
Gestational age[a]	0.83
Weight gain during pregnancy	0.76
Smoking during pregnancy[b]	0.83
Smoking before pregnancy[b]	0.82
Absolute alcohol per day during pregnancy[b,c]	0.53
Absolute alcohol per day prior to pregnancy[b,c]	0.60
PCB-contaminated fish consumption[b,d]	0.68

Values are Pearson correlation coefficients. All are significant at $p < 0.001$. $n = 236$. [a]Based on expected date of confinement. [b]Log transformed. [c]Estimated according to the method of Kuzma and Kissinger (68). [d]Based on an average annual sum in which each species of fish was weighted to reflect its degree of contamination.

findings from a single study should be treated as preliminary until replicated.

Retrospective Assessment

As noted earlier, prospective, longitudinal studies have several advantages over cross-sectional studies, including more accurate assessment of degree and timing of exposure and of relevant control variables. Given the high cost and complexity of longitudinal studies, however, some evidence of teratogenicity should be obtained retrospectively, if possible, before a full-scale prospective investigation is undertaken. Cross-sectional pilot studies focusing on highly exposed individuals can be valuable for identifying the most salient domains of impairment so that prospectively administered test batteries can be designed to focus on them. For example, attention deficits were first identified retrospectively in normal intelligence children of mothers known to have drunk alcohol during pregnancy on the basis of school records describing the children as hyperactive, easily distractible, and having a short attention span (65,66). Although the absence of prospective ascertainment of exposure makes the findings necessarily tentative, confirmation can subsequently be sought in a prospective study.

In our PCB research, certain control variable data obtained initially at delivery were obtained again at 4 years postpartum (9). As indicated in Table 7, the long-term reliability of maternal recall varied considerably depending on the domain being assessed. Mothers were remarkably accurate in recalling the birth weight of the child, reasonably reliable regarding gestational age, and somewhat less so about how much weight they had gained during pregnancy. Maternal report of smoking was markedly more reliable than for alcohol consumption, presumably because smoking is more habitual and, therefore, easier to recall. Validity coefficients for retrospective recall of drinking during pregnancy are also much weaker than for concurrent maternal report (67).

The recall coefficient for contaminated fish consumption before and during pregnancy (Table 7) was impressive given that fish consumption is much less habitual than smoking. This reliability is probably attributable to the fact that consumption of fresh-caught Lake Michigan fish, not available for purchase at the time, was a salient event for these families. The correlations of contaminated fish consumption with maternal serum and milk PCB levels were virtually the same for the reports obtained at delivery and 4 years later (r values = 0.34 and 0.37 for serum; 0.34 and 0.32 for milk), suggesting that the 4-year retrospective report may have been as valid as the report obtained at delivery. Thus, many important variables can be reliably assessed retrospectively, making it feasible in many cases to conduct one-shot, cross-sectional studies to guide the design and focus of more comprehensive prospective, longitudinal investigations and to supplement what is learned from them.

REFERENCES

1. Butcher RE. Behavioral testing as a method for assessing risk. Environ Health Perspect 18:75–78 (1976).
2. Weiss B, Spyker JM. Behavioral implications of prenatal and early postnatal exposure to chemical pollutants. Pediatrics 53:851–856 (1976).
3. Vorhees CV. Principles of behavioral teratology. In: Handbook of Behavioral Teratology (Riley EP, Vorhees CV, eds). New York:Plenum Press, 1986;23–48.
4. Spyker JM. Assessing the impact of low level chemicals on development: behavioral and latent effects. Fed Proc 34:1835–1844 (1975).
5. Riley EP. The long-term behavioral effects of prenatal alcohol exposure in rats. Alc: Clin Exp Res 14:670–673 (1990).
6. Needleman HL, Gunnoe C, Leviton A, Reed R, Peresie H, Maher C, Barrett P. Deficits in psychologic and classroom performance of children with elevated dentine lead levels. N Engl J Med 300:689–695 (1979).
7. Cohen J. Statistical Power Analysis for the Behavioral Sciences, 2d ed. Hillsdale, NJ:Erlbaum, 1988.
8. Schlesselman J. Case–Control Studies: Design, Conduct, Analysis. New York:Oxford University Press, 1982.
9. Jacobson SW, Jacobson JL. Early exposure to PCBs and other suspected teratogens: assessment of confounding. In: Longitudinal Studies of Infants Born at Psychological Risk (Greenbaum C, Auerbach J, eds). Norwood, NJ:Ablex, 1992;135–154.
10. Molfese V, Thomson B, Beadnell B, Bricker M, Manion L. Perinatal risk screening and infant outcome. Can predictions be improved with composite scales? J Reprod Med 32:569–576 (1987).
11. Caldwell BM, Bradley RH. Home Observation for Measurement of the Environment. Little Rock, AR:University

of Arkansas Press, 1979.

12. Barnard KE, Hammond MA, Booth CL, Bee HL, Mitchell SK, Spieker SJ. Measurement and meaning of parent-child interaction. In: Applied Developmental Psychology. Vol 3 (Morrison FJ, Lord C, Keating D, eds). New York:Academic Press, 1989;39–80.

13. Dunn LM, Dunn LM. PPVT Manual for Forms L and M. Circle Pines, MN:American Guidance Service, 1981.

14. Holmes TH, Rahe RH. The Social Readjustment Rating Scale. J Psychosomatic Res 11:213–218 (1967).

15. Crnic KA, Greenberg MT, Ragozin AS, Robinson NM, Basham RB. Effects of stress and social support on mothers and premature and full-term infants. Child Dev 54:209–217 (1983).

16. Beck AT, Ward CH, Mendelson M, Mock F, Erbaugh J. An inventory for measuring depression. Arch Gen Psych 4:561–571 (1961).

17. Derogatis LR. SCL-90R: Administration, Scoring, and Procedures Manual-II. Towson, MD:Clinical Psychometric Research, 1992.

18. Hyler SE, Skodol AE, Oldham JM, Kellman HD, Doidge N. Validity of the Personality Diagnostic Questionnaire-Revised: a replication in an outpatient sample. Compr Psychiatry 33:73–77 (1992).

19. Moos RH, Moos BS. Family Environment Scale Manual, 2d ed. Palo Alto, CA:Consulting Psychologists Press, 1986.

20. Straus MA. Measuring intrafamily conflict and violence: the Conflict Tactics (CT) Scales. J Marriage Fam 41:75–88 (1979).

21. Jacobson SW, Fein GG, Jacobson JL, Schwartz PM, Dowler JK. The effect of intrauterine PCB exposure on visual recognition memory. Child Dev 56:853–860 (1985).

22. Bellinger DC, Sloman J, Leviton A, Rabinowitz M, Needleman HL, Waternaux C. Low-level lead exposure and children's cognitive function in the preschool years. Pediatrics 87:219–227 (1991).

23. Dietrich KN, Berger OG, Succop PA, Hammond PB, Bornschein RL. The developmental consequences of low to moderate prenatal and postnatal lead exposure: intellectual attainment in the Cincinnati Lead Study cohort following school entry. Neurotoxicol Teratol 13:37–44 (1993).

24. Hans SL, Henson LG, Jeremy RJ. The development of infants exposed in utero to opioid drugs. In: Longitudinal Studies of Children at Psychological Risk: Cross-National Perspectives (Greenbaum CW, Auerbach JG, eds). Norwood, NJ:Ablex, 1992;155–173.

25. Garmezy N. Stress, competence, and development. Am J Orthopsychiatry 57:159–174 (1987).

26. Rutter M. Psychosocial resilience and protective mechanisms. In: Risk and Protective Factors in the Development of Psychopathology (Rolf J, Masten AS, Cicchetti D, Nuechterlein KH, Weintraub S, eds). Cambridge, UK:Cambridge University Press, 1990.

27. Jacobson JL, Jacobson SW. Prenatal alcohol exposure and neurobehavioral development: Where is the threshold? Alcohol Health Res World 18:30–36 (1994).

28. Cahalan D, Cisin IH, Crossley HM. American drinking practices: a national study of drinking behavior and attitudes. Monograph No 6. New Brunswick, NJ:Rutgers Center of Alcohol Studies, 1969.

29. Bowman RS, Stein LI, Newton JR. Measurement and interpretation of drinking behavior. Q J Stud Alcohol 36:1154–1172 (1975).

30. Jacobson JL, Jacobson SW, Sokol RJ, Martier SS, Ager JW, Kaplan-Estrin MG. Teratogenic effects of alcohol on infant development. Alcohol Clin Exp Res 17:174–183 (1993).

31. Jacobson JL, Jacobson SW. Methodological considerations in behavioral toxicology of infants and children. Dev Psychol (in press).

32. Zuckerman B, Frank DA, Hingson R, Amaro H, Levenson SM, Kayne H, Parker S, Vinci R, Aboagye K, Fried LE, Cabral H, Timperi R, Bauchner H. Effects of maternal marijuana and cocaine use on fetal growth. N Engl J Med 320:762–768 (1989).

33. Ambre J. The urinary excretion of cocaine and metabolites in humans. J Anal Toxicol 9:241–245 (1985).

34. Ostrea EM, Brady M, Gause S, Raymundo AL, Stevens M. Drug screening of newborns by meconium analysis. A large-scale, prospective, epidemiologic study. Pediatrics 89:107–113 (1992).

35. Miller V, Holzel A. Growth and development of endodermal structures. In: Scientific Foundations in Pediatrics (Davis JA, Dobbing J, eds). Philadelphia:WB Saunders, 1974;281–296.

36. Day N, Robles N. Methodological issues in the measurement of substance abuse. Ann NY Acad Sci 562:8–13 (1989).

37. Sameroff AJ, Seifer E. Familial risk and child competence. Child Dev 54:1254–1268 (1983).

38. Bradley RH, Caldwell BM, Rock SL, Ramey CT, Barnard KE, Gray C, Hammond MA, Mitchell S, Gottfried AW, Siegel L, Johnson DL. Home environment and cognitive development in the first 3 years of life: a collaborative study involving six sites and three ethnic groups in North America. Dev Psychol 25:217–235 (1989).

39. Gottfried AW, Gottfried AE. Home environment and cognitive development in young children of middle-socioeconomic-status families. In: Home Environment and Early Cognitive Development: Longitudinal Research (Gottfried AW, ed). Orlando, FL:Academic Press, 1984.

40. Jacobson SW, Jacobson JL. Breastfeeding and intelligence. Lancet 339:926 (1992).

41. Jacobson JL, Jacobson SW. Detecting the effects of prenatal drug exposure in socioenvironmentally-deprived children. In: Cocaine Mothers and Cocaine Babies: The Role of Toxins in Development (Lewis M, Bendersky M, eds). Hillsdale, NJ: Erlbaum, 1995;111–127.

42. Jacobson SW, Jacobson JL, Frye KF. Incidence and correlates of breast-feeding in disadvantaged women. Pediatrics 88:728–736 (1991).

43. Alessandri SM, Sullivan MW, Imaizumi S, Lewis M. Learning and emotional responsivity in cocaine-exposed infants. Dev Psychol 29:989–997 (1993).

44. Streissguth AP, Aase JM, Clarren SK, Randels SP, LaDue RA, Smith DF. Fetal alcohol syndrome in adolescents and adults. JAMA 265:1961–1967 (1991).

45. Ernhart CB. Cofactors in observational research: issues and examples from the lead effects literature. In: Behavioral Toxicology of Childhood (Melton GR, Schroeder SR, Sonderegger TB, eds). Lincoln, NB:University of Nebraska Press, in press.

46. Tabachnick BG, Fidell LS. Using Multivariate Statistics. New York:Harper & Row, 1983.

47. Kleinbaum DG, Kupper LL, Muller KE. Applied Regression Analysis and Other Mutivariable Methods 2d ed. Boston:PWS-Kent, 1988.

48. Jacobson JL, Jacobson SW, Humphrey HEB. Effects of in utero exposure to polychlorinated biphenyls and related contaminants on cognitive functioning in young children. J Pediatr 116:38–45 (1990).

49. Jacobson SW, Jacobson JL, O'Neill JM, Padgett RJ, Frankowski JJ, Bihun JT. Visual expectation and dimensions of infant information processing. Child Dev 63:711–724 (1992).

50. Day NL, Richardson G, Robles N, Sambamoorthi U, Taylor P, Scher M, Stoffer D, Jasperse D, Cornelius M. Effect of prenatal alcohol exposure on growth and morphology of offspring at 8 months of age. Pediatrics 85:748–752 (1990).

51. Dietrich KN, Succop PA, Berger OG, Hammond PB, Bornschein RL. Lead exposure and the cognitive development of urban preschool children. Neurotoxicol Teratol 13:303–311 (1991).

52. Fried PA, Watkinson A. 12- and 24-month neurobehavioural follow-up of children prenatally exposed to marijuana, cigarettes and alcohol. Neurotoxicol Teratol 10:305–313 (1988).

53. Ryan L, Ehrlich S, Finnegan L. Cocaine abuse in pregnancy: effects on the fetus and newborn. Neurotoxicol Teratol

9:295–299 (1987).

54. Frank DA, Bauchner H, Parker S, Huber AM, Kyel-Aboagye K, Cabral H, Zuckerman B. Neonatal body proportionality and body composition after *in utero* exposure to cocaine and marijuana. J Pediatr 117:622–626 (1990).

55. Woods JR, Plessinger MA, Clark KE. Effect of cocaine on uterine blood flow and fetal oxygenation. JAMA 257:957–961 (1987).

56. O'Connor MJ, Sigman M, Kasari C. Interactional model for the association among maternal alcohol use, mother-infant interaction, and infant cognitive development. Inf Behav Dev 16:177–192 (1993).

57. Biddle BJ, Marlin MM. Causality, confirmation, credulity, and structural equation modeling. Child Dev 58:4–17 (1987).

58. Sokol RJ, Ager J, Martier S, Debanne S, Ernhart C, Kuzma J, Miller SI. Significant determinants of susceptibility to alcohol teratogenicity. Ann NY Acad Sci 477:87–100 (1986).

59. Rutter M, Quinton D. Long-term follow-up of women institutionalized in childhood: factors promoting good functioning in adult life. Br J Dev Psychol 18:225–234 (1984).

60. Streissguth AP, Barr HM, Martin DC, Herman CS. Effects of maternal alcohol, nicotine, and caffeine use during pregnancy on infant mental and motor development at 8 months. Alcohol Clin Exp Res 4:152–164 (1980).

61. Greene T, Ernhart CB, Ager J, Sokol R, Martier S, Boyd T. Prenatal alcohol exposure and cognitive development in the preschool years. Neurotoxicol Teratol 13:57–68 (1991).

62. Richardson GA, Day NL. Maternal and neonatal effects of moderate cocaine use during pregnancy. Neurotoxicol Teratol 13:455–460 (1991).

63. McCall RB. Nature-nurture and the two models of development: a proposed integration with respect to mental development. Child Dev 52:1–12 (1981).

64. Jacobson SW, Jacobson JL, Sokol RJ. Effects of fetal alcohol exposure on infant reaction time. Alc: Clin Exp Res 18:1125–1132 (1994).

65. Shaywitz SE, Cohen DJ, Shaywitz BA. Behavior and learning difficulties in children of normal intelligence born to alcoholic mothers. J Pediatr 96:978–982 (1980).

66. Aronson M, Kyllerman M, Sabel K-G, Sandin B, Olegard R. Children of alcoholic mothers: developmental, perceptual and behavioral characteristics as compared to matched controls. Acta Paediatr Scandinav 74:27–35 (1985).

67. Jacobson SW, Jacobson JL, Sokol RJ, Martier SS, Ager JW, Kaplan MG. Maternal recall of alcohol, cocaine, and marijuana use during pregnancy. Neurotoxicol Teratol 13:535–540 (1991).

68. Kuzma JW, Kissinger DG. Patterns of alcohol and cigarette use in pregnancy. Neurobehav Toxicol Teratol 3:211–221 (1981).